Statistical Analysis Of Variant Experiments

1 Introduction

Controlled experiments are different from observational studies, one typically hears about in the news. In an observational study, the subjects themselves choose their assignment to the different experiment groups, while the investigator interprets the results. By contrast, in a controlled experiment, the investigator gets to decide who will be in the treatment group and who will be in the control.

Thus, studies on the effects of smoking are necessarily observational: no one will start smoking because the investigator wants him to. Nonetheless, there is a vast body of research, comparing smokers to non-smokers as if they are treatment and control groups. Invariably, the smokers come off badly in these comparisons and many social policies have been enacted based on these studies, even though what they really show is an association, which is only a circumstantial evidence for causation. In fact, the storied British statistician R. A. Fisher did not believe in statistical evidence against cigarettes and suggested possible confounding variables. (They were shown to be implausible by later studies.)[1]

In a *randomized controlled experiment* (R.C.E.), the investigator assigns subjects to the different groups randomly, thus, on average, removing any possibility that the difference between the treatment and the control could be explained by anything other than the treatment itself. Thanks to this quality, R.C.E.s are considered the gold standard of experiment design and, lucky for us, can be employed in the investigation of differences in the behavior of interactive computer applications.

In the next chapter we will review the basics of statistical hypothesis testing—the field of mathematical statistics, concerned with analyzing controlled experiments. The fundamental idea there is to develop a procedure that will enable the researcher to make a claim about the entire population with a given degree of certainty, based on a set of sample observations. Specifically, we want to be able to detect the moment when the difference in the value of some metric, as observed in the samples, is less likely due to chance alone than a given (small) probability, α referred to as the test's *significance level*.

In chapter 3 we will address the question of practical applicability of the methods developed in chapter 2 for analyzing Variant experiments. In particular, we will develop an approach to two

practical problems: how to avoid calling an experiment too soon due to spurious significance, and how to avoid running an experiment for too long, chasing a difference that may not be there.

2 Z-test Of Statistical Significance

2.1 Statistical Hypotheses

The central idiom of inferential statistics is the *null hypothesis* — a statement of status quo, or of no difference between the control and the treatment populations. As applied to testing of interactive computer application, the null hypothesis may state that the difference in user subscription rate in the new (treatment) experience is the same as in the existing (control) experience. The opposite of the null hypothesis is known as the *alternative hypothesis*, where the researcher conjectures a difference between the control and the treatment populations.

It is intuitively obvious that the larger the number of sample observations, the more likely any observed difference correctly predicts a similar difference in the entire population. However, the researcher is limited in time and, possibly, other resources, and wants to call an experiment on as small a sample size as possible, thus risking making a mistake. The goal of the Z-test procedure is to establish as small a sample size as possible on which a systematic inference can be made.

There are two different mistakes the researcher can make: *Type I error*, or false positive, is when, on the basis of the observed sample observations, the researcher concludes that there's a difference between the two populations, when, in face, there isn't one. *Type II error*, or false negative, arises when the researcher concludes that there isn't a difference between the two populations, when, in fact, there is one.

Type I error is more critical than type II because it may lead to replacing an existing experience with one that is worse. Type II error is less critical (we may miss what would have been a small improvement) but is progressively harder to avoid, as the difference we are trying to detect becomes smaller.

The probability of committing type I error is traditionally denoted α and is referred to as the *significance level* of the test. Typically, it is set at .10 or, for critical applications, .05. The closer it is to 0, the more confident the researcher can be about his claim, paying for this confidence with a larger number of observations he will have to make to get there.

2.2 Comparing Two Proportions

We start with the case when the measure of interest is a proportion, e.g. a conversion rate from one Web page to the next. For the given sample sizes n_1 and n_2, we can calculate cumulative

observed proportions p_1 and p_2, both of which are in the range [0,1], as the sum of all observations which resulted in the desirable outcome, divided by the sum of all observations.

The standard statistical procedure for determining this is the two-tailed z-test for *statistical significance*, according to which we look for the smallest sample sizes n_1 and n_2, which, given the significance level α, result in the probability of the event that the difference $p_2 - p_1$ is greater than some value δ be less than α, i.e.

$$P\left(\left|p_2 - p_1\right| > \delta\right) \leq \alpha$$

Assuming the binomial distribution of both ratios, we can compute the *confidence interval* (CI) around the difference, which determines whether or not the observed difference is more or less likely than α:

$$\delta^{\pm} = p_2 - p_1 \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1\left(1-p_1\right)}{n_1^{\,2}} + \frac{p_2\left(1-p_2\right)}{n_2^{\,2}}} \qquad (1)$$

where $z_{\alpha/2}$ is the z-score, corresponding to the desired significance level α divided by 2 to account for the fact that we don't know ahead of time if the difference is positive or negative. Z-scores are found in standard distribution tables, or computed with standard statistical functions in most advanced data analysis tools, such as Excel.

If the low confidence limit δ^- is greater than zero, then we claim that the observed ratio p_2 is statistically significantly higher than p_1 with the significance level α. Conversely, if the upper confidence limit δ^+ is less than zero then we conclude that the observed ratio p_2 is statistically significantly lower than p_1 with the significance level α.

Note that we don't make any claims about the magnitude of the difference between p_2 and p_1 in the entire population — only that it exists.

2.3 Comparing Two Averages

Let's now turn to the case when we're interested in comparing not proportions, but averages of continuous values, such as a monetary amount or number of units sold. For the given sample sizes n_1 and n_2, we can calculate cumulative observed means μ_1 and μ_2 as

$$\mu_1 = \sum_{i=1}^{n_1} x_{1_i} \text{ and } \mu_2 = \sum_{j=1}^{n_2} x_{2_j}$$

Just as we did with proportions, we compute the confidence interval (CI) around the difference μ_2 - μ_1 , which, for this case looks like this:

$$\delta^{\pm} = \mu_2 - \mu_1 \pm z_{\alpha/2} \cdot \sqrt{\frac{\sum_i \left(x_{1i} - \mu_1\right)^2}{n_1^2} + \frac{\sum_j \left(x_{2j} - \mu_2\right)^2}{n_2^2}} \qquad (2)$$

As before, if the low confidence limit δ^- is greater than zero, we claim that $\mu_2 > \mu_1$ (or, conversely, if the upper confidence limit δ^+ is less than zero, that $\mu_2 < \mu_1$) with the significance level α.

3 Practical Analysis Of Variant Experiments

3.1 When To Call an Experiment

The challenge of applying formulas (1,2) is that they do not directly address the question that the researcher really needs to answer: when to end an experiment. Is it as soon as it reaches statistical significance? And what if that significance is too slow to materialize, how long to wait for it before giving up?

In general, these three factors influence statistical significance:

- The size of the observed difference between the two groups
- The number of samples observed so far
- The significance level α we wish to attain

The larger the difference between the two metrics, the larger is the first term in (1,2), hence the fewer samples it takes to conclude significance. This makes sense: the difference in sampled values is the best evidence for a difference in the population values — this is the whole premise of what we're doing! Similarly, the greater the sample size, the smaller the second term in (1,2), causing the width of the confidence interval to shrink around the observed difference. Finally, the z function has the opposite effect: the larger the significance level, the larger z_α, the wider the confidence interval, the more samples it will take to narrow it down.

In order to confidently analyze Variant experiments, the researcher must plot both the metric of interest and the confidence interval, like in figure 1 below.

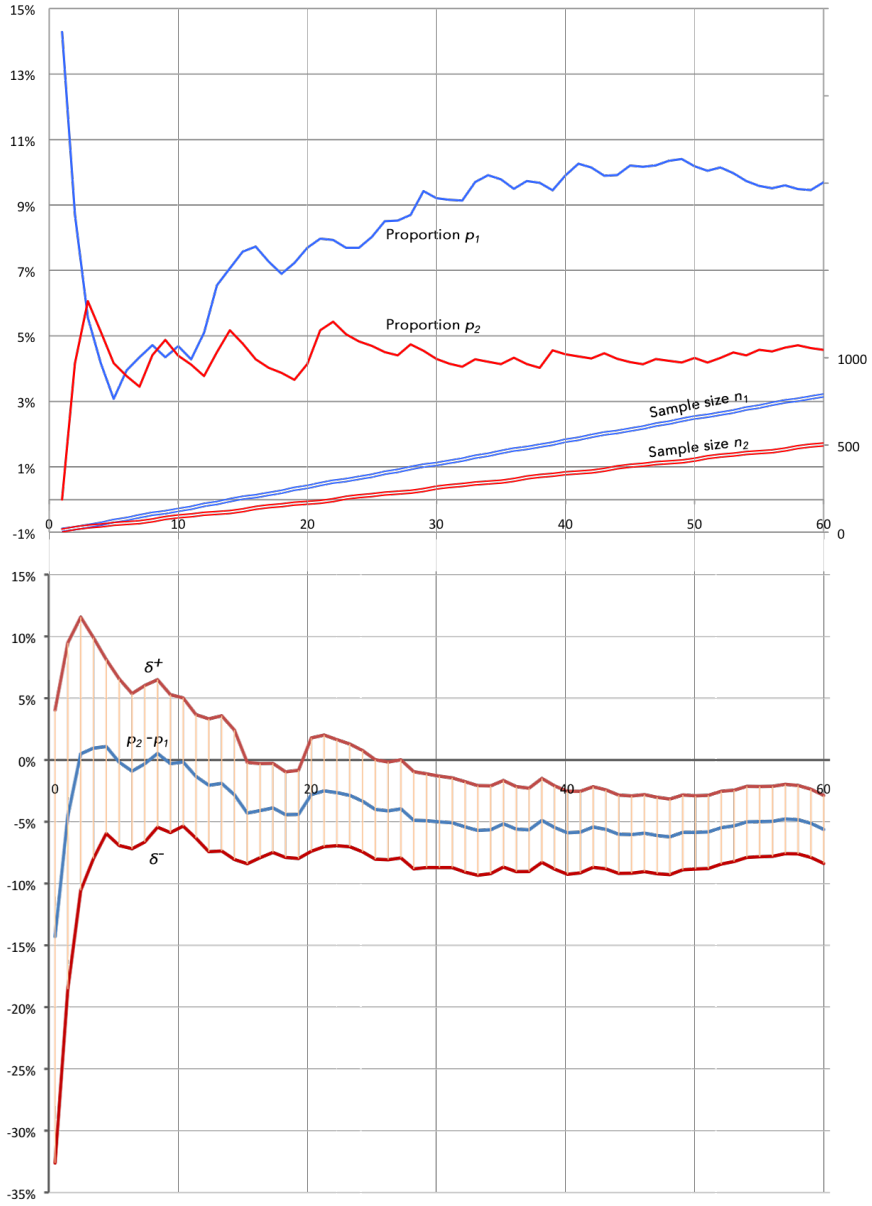

Figure 1. 95% confidence intervals around the difference of two cumulative proportions p_1 and p_2, aligned temporally with their values and sample sizes.

In the top graph, the two cumulative proportions p_1 and p_2 over time are plotted against the corresponding sample sizes n_1 and n_2, and the bottom graph presents the confidence interval $\delta^{\pm}$ around the difference between the two proportions, computed for the confidence level of 95%. What this figure really means is the following. Suppose we draw a vertical line through both graphs, say at the time unit of 30, corresponding to the sample sizes of $n_1 = 391$ and $n_2 = 256$. The corresponding values for the two proportions are $p_1 = 0.0942$ and $p_2 = 0.0455$, and for the

CI $\delta^{\pm} = [-0.0095, -0.0881]$. As discussed in the previous section, the fact that the upper bound of the CI lies below zero means that the experiment has reached statistical significance, i.e. we can be 95% confident that, taken over the entire population, p_1 is greater than p_2.

It is intuitive to interpret this result as the condition for calling the experiment. However, as a general rule, this would be wrong. Note how the upper confidence bound has already dipped briefly below zero around time unit 15, but came back up into the no confidence territory by time unit 20. What happened then is informally referred to as *spurious significance* and it is a common cause researchers conclude that the observed difference is significant when in fact it is not, i.e. commit the type I error.

In the next section we will consider a practical approach to how to avoid this kid of mistake, while Section 3.3 addresses the opposite problem of avoiding the type II errors.

3.2 Avoiding Type 1 Error

An important idea to understand about applying z-test to analyzing Variant R.C.E.s is that z-test does not consider variability of the sample values. But, precisely because of this variability, the confidence interval also varies and may come in and out of significance. In practice, this means that the researcher cannot call the experiment just as soon as it reaches significance.

Surprisingly, there is no systematic research on this topic. However, at Variant, we've developed a sound empirical method that we've calibrated with many customers and over many experiments: an experiment is truly significant when it not only reaches statistical significance, but also remains statistically significant for a period of time. The length of this period depends on the significance level α: for a 90% level tests, we recommend that the experiment continue for as long as 50% more traffic passes through it, as it took to get to significance, and for a 95% level test — 65% more traffic.

3.3 Avoiding Type 2 Error

Let's now consider the case when an experiment never seems to reach significance, as in figure 2 below.

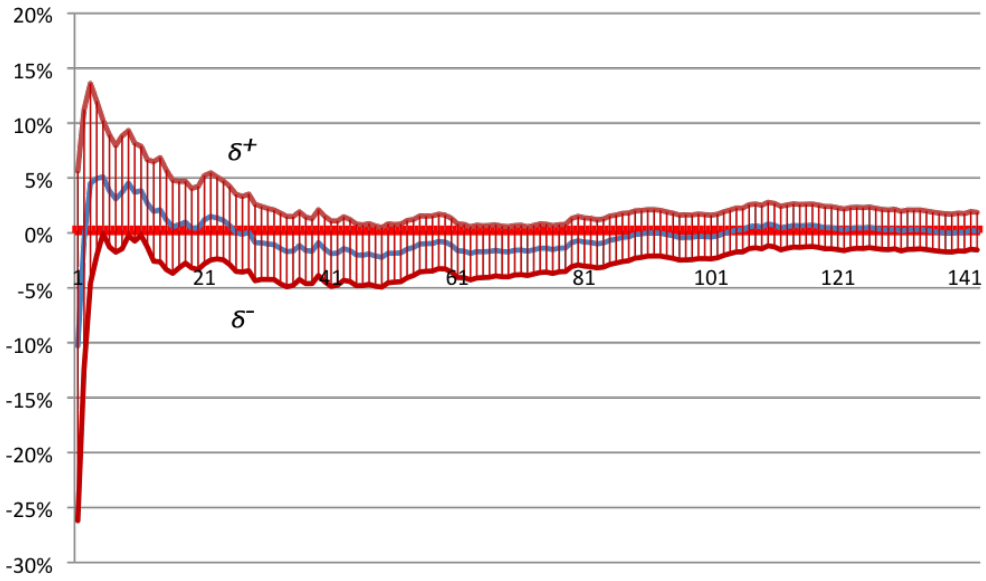

Figure 2. An inconclusive experiment: upper CI bound stays above 0 and lower CI bound stays below 0.

The problem here is that the upper CI bound δ^+ stays above zero and the lower CI bound δ^- — below. We may choose to continue to wait, hoping that the additional observations will yield a more precise experiment: indeed, as the sample sizes grow, the width of the ribbon diminishes, so there's a chance that one of the bounds might cross zero. Alternatively, we may try increasing the significance level α, if that's acceptable by the business requirements.

If increasing α is not an option, the solution is to continue the experiment until the CI ribbon becomes narrower than some threshold difference between p_1 and p_2 considered relevant by the business requirements. For example, in figure 2 we ended the experiment around time unit 141 when the width of the CI ribbon reached 5%. In many real life cases differences as small as 1% may be considered relevant, which may take several thousands of observations to reach.

3.4 Conclusion

To recap, a Variant experiment is callable when either of these two conditions has been met:

- If the experiment reaches statistical significance, i.e. when the CI ribbon entirely rises above zero or entirely falls below it, and remains significant for 50% more observations than it took to get to significance for 0.10 level tests (65% more observations than it took to get to significance for .05 level tests), the experiment is called by accepting the alternative hypothesis, or, in other words, the treatment wins.

- If the experiment does not reach statistical significance, while the CI ribbon has narrowed to where its width represents a difference between the treatment and the control that is not consequential to the application semantics, the experiment is called by rejecting the research hypothesis, or, in other words, the treatment fails to win and we stick with the control.

4 References

[1] David Freedman, Robert Pisani, and Roger Purves. Statistics. Norton, New York, 1978.

[2] Allan G. Bluman, Elementary Statistics, 2004 http://www.amazon.com/Elementary-Statistics-Step-Approach/dp/0072549076#reader_0072549076, Chapter 9 (http://faculty.ccc.edu/colleges/hwashington/math/Resources/Stats_PDF_Chapters/ch09.pdf)

[3] D.S. Moore and G. P. McCabe, Introduction to the Practice of Statistics, Freeman, 5th edition, 2006.